VALIDITAS TES DAN KUALITAS BUTIR SOAL

Oleh: Asyraf Muzaffar

Abstrak

Validitas tes adalah sifat yang sangat penting yang melekat pada sebuah tes karena dari seluruh proses pengukuran atribut psikologis, puncaknya ada pada masalah validitas. Dalam kerangka validitas tes, skor yang dihasilkan dari proses panjang perencanaan, pengembangan dan pelaksanaan tes ditafsirkan dan dijadikan dasar pengambilan keputusan dan tindakan.Berbagai macam bukti empiris terkait (construct, isi dan kriteria) dan dasar pikir teoritis yang melatarbelakangi keputusan dan tindakan tersebut dinilai kesesuaian dan kecukupannya. Demikian juga konsekwensi sosial dan implikasi nilai (value) dari penafsiran dan keputusan tersebut diintegrasikan ke dalam penilaian validitas tes.Karena itu agar penilaian atau studi validitas memiliki tingkat kesahihan yang tinggi, data atau skor tes yang dijadikan dasar kegiatan tersebut haruslah yang memiliki tingkat galat atau kesesatan pengukuran yang rendah. Skor tes yang demikian hanya dapat diperoleh jika butir-butir soal yang digunakan dalam tes tersebut memenuhi kriteria-kriteria butir soal yang baik.Diantara kriteria terpenting adalah adanya construct yang didefinisikan dan dipahami dengan jelas dan dipastikan kesesuaian yang tinggi setiap butir soal dengan construct tersebut. Kriteria lainnya adalah terpenuhinya asumsi teknis psikometri, yaitu unidimensionalitas butir soal dan local independence. Demikian juga butir soal yang baik hendaklah sejalan dengan etika dan hukum yang berlaku.

Kata kunci: Validitastes dan Butir Soal

A. Pendahuluan

Ketika menilai atau meneliti validitas sebuah tes, fokus perhatian akan diarahkan kepada kesimpulan dan penafsiran terhadap hasil (skor) tes tersebut dengan melihat kepada bukti-bukti validitas yang didapatkan. Dalam pandangan yang lebih mutakhir, penilaian validitas tes bahkan memasukkan penilaian pilihan

tindakan dan keputusan yang diambil, sejauh mana dan seperti apa konsekwensi sosial dan implikasi nilai (*value*) dari tindakan atau pengambilan keputusan tersebut. Pengambilan keputusan sendiri merupakan *raison d'tre* dari diadakan atau dilaksanakannya sebuah tes.

Dalam penilaian validitas yang berfokus pada kesimpulan dan penafsiran tersebut, kualitas tes sendiri tidak secara khusus menjadi perhatian, konon lagi kualitas dari butir-butir soal. Padahal butir-butir soal merupakan elemen-elemen dasar yang membentuk sebuah tes. M.T. Kane, seperti dikutip Haladyna, secara jelas menyatakan bahwavaliditas bukan sifat dari tes tetapi sifat dari penafsiran skor tes dan penggunaannya[1]. Dengan demikian istilah validitas tes sebenarnya bermakna validitas penafsiran hasil tes.

Dalam pembicaraan informal, orang kadang-kadang berbicara tentang validitas tesdengan maksud validitas tes itu sendiri; tapi tidak pernah ada yang menyebutkan validitas butir soal. Jika ada istilah atau konsep yang berbicara tentang validitas butir soal, maka konsep tersebut dapat dikatakan tidak bermakna karena cakupannya sangat sempit. Apa lagi, seperti yang dikatakan oleh Haladyna pengukuran dengan hanya satu butir soal tidak saja rentan dengan kesalahan kesimpulan, tetapi juga atribut manusia terlalu kompleks untuk diukur dengan hanya satu butir soal.[2]

Bahwa penilaian validitas tes lebih melihat kepada kesahihan kesimpulan yang ditarik berdasarkan bukti-bukti pendukung adalah sesuatu yang memiliki dasar yang kuat. Review sekilas terhadap definisi-definisi validitas tes dalam buku-buku literatur psikometri memperlihatkan secara dominan fokus penilaian tersebut. Tetapi perlu disadari bahwa kenyataan tersebut tidak boleh dipahami bahwa kualitas tes dan butir-butir soal tidak menentukan dalam penilaian validitas tes. Butir-butir soal merupakan bahan-bahan dasar yang secara bersama-sama membentuk sebuah tes dan, karena itu, kualitas butir-butir soal tersebut akan menentukan kualitas tes secara keseluruhan. Adalah tidak mungkin sebuah tes yang baik terdiri dari butir-butir soal yang tidak berkualitas baik. Dan selanjutnya, sulit untuk menarik

---

[1]Haladyna, Thomas M., *Developing and Validating Multiple-Choice Test Item* (2nd Ed), (New Jersey: Lawrence ErlbaumAssociates, 1999) hal. 9

[2]Ibid , hal. 4

kesimpulan dengan validitas yang memadai berdasarkan hasil sebuah tes dengan kualitas yang jelek.

Berdasarkan dasar pemikiran yang disampaikan di atas, maka sudah seharusnya kualitas butir soal mendapatkan perhatian yang tinggi dalam proses penulisan dan pengembangannya, walaupun kualitas butir soal tidak secara eksplisit disebutkan dalam penilaian validitas. Untuk menghasilkan butir soal yang berkualitas, ada banyak hal yang harus diperhatikan, tahapan proses dan prosedur yang mesti dilalui dan pilihan, kesulitan serta tantangan yang harus dipertimbangkan dengan matang untuk sampai kepada keputusan atau pilihan yang tepat. Jika semua proses tersebut telah dilalui dengan benar dan menghasilkan butir-butir soal berkualitas dalam sebuah tes yang baik, maka tepatlah penilaian validitas akan secara langsung fokus kepada kesimpulan, penafsiran dan bahkan tindakan yang diambil berdasarkan hasil dan bukti yang diperoleh dari sebuah tes dan studi validasi.

B. Validitas : Konsep dan Konsekwensi

1. Konsep Validitas

Pengertian validitas sebagai sebuah konsep yang sangat penting dalam pengukuran mental telah berkembang sedemikian rupa seiring perkembangan disiplin ilmu psikometri, konsep pendidikan dan bahkan kehidupan secara keseluruhan. Namun demikian belum terlihat adanya kesepakatan yang kuat di antara para ahli tentang sebuah pengertian validitas yang dapat diterima secara mayoritas. Masih terlihat adanya "aliran" yang berbeda-beda dalam memahami konsep yang sangat mendasar dalam disiplin ilmu pengukuran mental ini.

Osterlind telah merangkum perkembangan pengertian validitas sejak masa awal konsep ini dikemukakan hingga akhir abad kedua puluh yang lalu.[3] Uraian sejarah perkembangan validitas berikut ini, didasarkan pada rangkuman Osterlind.

Salah satu pengertian validitas yang dikemukakan pada awal-awal perkembangan disiplin psikometri adalah oleh Garrett. Garrett pada tahun 1937 mendeskripsikan validitas sebagai

---

[3] Osterlind,Steven J., *Constructing Test Items: Multiple-Choice, Constructed Response, Performance and Other Formats*, (2nd Ed), (Boston: Kluwer Academic Publishers, 1998) hal. 61

ketepatan sebuah tes mengukur apa yang dimaksudkan untuk diukur oleh tes tersebut. Selanjutnya lebih dari tiga dasawarsa kemudian, yaitu tepatnya pada tahun 1971, Cronbach mengemukan pandangan tentang makna validitas yang fokusnya berbeda dengan Garrett. Bagi Cronbach validitas tes ditegakkan melalui sebuah proses validasi dimana bukti dikumpulkan oleh pengembang tes untuk mendukung jenis-jenis penafsiran yang sesuai, yang disimpulkan dari skor-skor tes. Bila kita membandingkan kedua definisi ini terlihat jelas bahwa Cronbach lebih menekankan pada penafsiran skor tes, bukan pada tes itu sendiri.

Pada tahun-tahun yang lebih akhir , pengertian validitas masih saja bervariasi pada penafsiran, objek pengukuran dan instrumen alat ukur itu sendiri. Mehrens dan Lehmann (1987) menganggap validitas paling tepat didefinisikan sebagai sejauh mana penafsiran tertentu dapat dibuat secara akurat – dan suatu tindakan didasarkan padanya - dari skor tes dan pengukuran lainnya. Sementara itu dalam pandangan Anastasi (1997) validitas terkait dengan apa yang diukur oleh tes dan seberapa baik tes tersebut mengukurnya.

Tetapi ada satu hal yang penting untuk digarisbawahi dari pengertian validitas pada periode tersebut bahwa pandangan tentang validitas mengalami penambahan aspek penekanan dalam sebuah pemikiran yang dikemukakan oleh Samuel Messick. Samuel Messick mendefinisikan validitas sebagai "… *an overall evaluative judgement of the degree to which empirical evidence and theoritical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment.*"[4]Dalam sumber yang lain Messick menyampaikan pandangan yang serupa, yaitu "*Validity is an overall evaluative judgment, founded on empirical evidence dan theoritical rationales, of the adequacy dan appropriateness of inferences and actions based on test scores*."[5]Dalam pengertian ini, validitas tes tidak hanya mempersoalkan penafsiran terhadap hasil tes beserta bukti empiris yang mendukungnya dan dasar pikir teoritis yang menjadi fondasi penafsiran tersebut, tetapi juga

[4]Messick, S., "Validity", dalam Linn,R. L., *Educational Measurement* (3rd Ed), (New York: Macmillan, 1988), hal. 14

[5]Osterlind,Steven J., *Constructing Test …*hal. 62

tindakan yang didasarkan pada penafsiran tersebut. Dengan demikian Messick memasukkan konsekwensi dari tindakan atau keputusan yang didasarkan pada penfsiran tes ke dalam penilaian validitas.

Pandangan Messick tentang validitas ini menemukan relevansinya dalam konteks kehidupan sekarang. Sudah menjadi pengetahuan umum, setidaknya sebagian sekolah melakukan tindakan-tindakan tertentu seperti memotivasi siswa secara berlebihan, menghapus jawaban salah dan mengganti dengan yang benar pada lembar jawaban atau siswa-siswa dibiarkan secara massif dan massal melakukan kecurangan untuk meningkatkan skor tes dalam ujian-ujian penting seperti ujian nasional. Tindakan-tindakan tersebut dilakukan, baik karena tekanan eksternal ataupun karena motivasi dan kepentingan internal, tidak berangkat dari kepentingan untuk memperbaiki proses belajar siswa atau untuk meningkatkan mutu capaian hasil belajar. Pada titik ini persoalan validitas tes tidak lagi eksklusif sebagai masalah tes dan penafsirannya, tetapi telah bersinggungan atau bahkan berhimpitan dengan persoalan kehidupan yang lebih luas.

2. Kesatuan Konsep Validitas

Sebelum membahas konsekwensi dari penafsiran dan tindakan atau keputusan yang didasarkan pada sebuah tes, penulis terlebih dahulu perlu membahas validitas sebagai sebuah kesatuan konsep (*unitary concept*). Yang dimaksud dengan kesatuan konsep adalah bahwa validitas tidak dapat dibagi-bagi ke dalam jenis-jenis validitas. Pandangan lama adanya jenis-jenis validitas yang terbagi kepada*validitas construct*, *validitas isi* dan*validitas terkait kriteria*, sesungguhnya lebih mencerminkan jenis-jenis bukti pendukung dalam menilai validitas tes. Jenis-jenis bukti tersebut secara bersama-sama akan menentukan derajat atau tingkat validitas penafsiran hasil sebuah tes. Dalam pemahaman baru ini, maka penyebutan yang lebih tepat adalah *bukti validitas yang terkait construct*, *bukti validitas yang terkait isi* dan *bukti validitas terkait kriteria*.[6] Adapun validitas tes itu sendiri adalah satu kesatuan.

Penilaian validitas bukti-bukti tersebut dilakukan melalui penelitian atau studi validitas dengan prosedur yang lazim.

[6]Ibid., hal.64

Validitas isi dinilai dengan meneliti representasi isi tes terhadap domain atau konstruk yang diukur oleh tes tersebut. Validitas terkait kriteria (validitas sama saat dan validitas ramalan) diteliti dengan melihat korelasi skor-skor dalam sebuah tes dengan kriteria kinerja dalam bidang berkaitan pada masa yang sama atau masa yang berbeda. Dan validitas construct dinilai berdasarkan sejauh mana bukti-bukti yang tersedia mendemonstrasikan hubungan logis dan matematis antara construct dan tes yang dimaksudkan untuk mengukur construct tersebut seperti yang dihipotesakan sebelumnya.[7]

Disamping pemahaman validitas sebagai kesatuan konsep, penting juga untuk diperhatikan bahwa validitas tes sering kali tidak dapat dinyatakan dalam dikotomi valid – tidak valid. Penilaian validitas lebih cenderung mengambil deskripsi dalam derajat. Jika sebuah penafsiran didukung oleh bukti yang banyak dan kuat, maka penafsiran tes dapat disimpulkan memiliki derajat validitas yang tinggi. Sebaliknya, penafsiran yang didukung oleh sedikit bukti atau bukti yang lemah, maka derajat validitasnya bisa disimpulkan rendah.[8]

3. Konsekwensi Validitas Tes

Sebagai instrumen pengukuran atribut psikologis semua tes pada akhirnya akan sampai pada suatu tahap proses dimana data-data dan informasi-informasi yang dikumpulkan melalui tes tersebutakan ditafsirkan untukmemaknai dan menyimpulkan derajat dan kualitas terkait atribut psikologis yang diukur oleh instrumen tersebut. Hasil dari penafsiran, pemaknaan dan penyimpulan tersebut padatahap selanjutnya akan menjadi basis bagi kebijakan dan tindakan evaluatif dan korektif yang diambil oleh pihak-pihak terkait yang berwenang.Sering kali kebijakan dan tindakan evaluatif dan korektif itu memberikan dampak positif dan konstruktif atau negatif dan destruktif yang signifikan bagi individu atau kelompok yang menjadi subjek dari sebuah tes. Bahkan acap kali tindakan berbasis hasil tes itu akan menentukan kualitas hidup

---

[7]Untuk lebih mendalami konsep validitas, dapat dibaca Crocker,Linda& Algina, James,Introduction to Classical and Modern Test Theory, (New York: CBS College Publishing, 1986), hal. 217 - 239

[8]Osterlind,Steven J., *Constructing Test ...,* hal.64

atau karir seseorangatausekelompok orang sepanjang hayat mereka, baik secara positif maupun negatif.Dengan konsekwesi yang sangat menentukan ini, maka sudah seharusnyalah individu-individu yang menjadi subjek tes tersebut mendapat hasil yang adil. Kondisi ini hanya bisa dicapai jika penilaian dan penafsiran hasil tes memiliki validitas yang memadai, yang didasarkan pada bukti-bukti yang sahih.

Disamping implikasi tes yang berpengaruh terhadap individu, hasil tes juga dapat berpengaruh terhadap program, projek dan suatu materi pendidikan atau pembelajaran. Tidak jarang pengambilan keputusan tentang program, projek dan materi pembelajaran didasarkan pada penafsiran hasi tes dengan validitas rendah sehingga menyebabkan pilihan tindakan yang keliru.

Dalam kenyataannya banyak praktisi dan pemangku kepentingan (*stake holder*) pendidikan seperti pengambil kebijakan di bidang pendidikan, guru dan pengawas terlihatkurang memberikan perhatian dan kurang menyadari implikasi krusial sebuah tes atau rangkaian tes.Padahaltes yang bahkan kelihatan kurang penting sekalipunmerupakan instrumen yang memiliki implikasi dan konsekwensi konstruktif dan/atau destruktif yang substansial bagi individu dan kelompok yang kepada mereka diterapkan instrumen tersebut.Kadang-kadang implikasi itu tidak mengejawantah dalam jangka pendek, tetapi baru terlihat jauh di kemudian hari. Oleh karena signifikannya akibat yang ditimbulkan oleh tes terhadap hidup individu dan kelompok, maka sudah seharusnya semua aspek dan tahap terkait pengembangan dan pelaksanaan tes dilakukan secara teliti dan dipastikan bahwa hal-hal yang diperlukan untuk menjamin kualitas tes yang baik telah dilakukan. Demikian juga penafsiran, pemaknaan dan penyimpulan hasil tes hendaklahdipertanyakan telah sejauh mana melalui kegiatan-kegiatan standaruntuk menyediakan bukti-bukti yang mendukung penyimpulan yang dikemukakan sehingga validitas yang memadai untuk penyimpulan tersebut dapat dipertanggungjawabkan.

Butir soal dan tes yang berkualitas baik memang belum menjamin validitas penafsiran dan penyimpulan hasil tes dengan derajat yang tinggi. Hal tersebut disebabkan oleh karena validitas penafsiran hasil tes merupakan proses tersendiri yang bisa saja dipengaruhi oleh faktor-faktor yang lain. Sebagai hasil sebuah

penafsiran, validitas tes dapat dipengaruhi oleh subjektifitas orang yang melakukan penilaian validitas, baik disengaja atau tidak, dan oleh karenanya sangat memungkinkan terjadinya perbedaan pendapat diantara orang-orang yang terlibat. Bahkan salah satu kritikan dan tuduhan yang dialamatkan kepada tes adalah tentang adanya sebagian hasil tes yang ditengarai telah disalahtafsirkan dan disalahgunakan.[9]

Namun demikian butir soal dan tes yang berkualitas baik akan menjadi basis yang kuat bagi penafsiran dan penyimpulan hasil tes dengan kesesatan (galat) pengukuran minimal dan dapat membantu meminimalisir perbedaan penafsiran. Merupakan suatu kesulitan yang luar biasa, jika tidak dikatakan mustahil, penafsiran dan penyimpulan hasil sebuah tes akan mencapai validitas yang memadai dan dapat dipertanggungjawabkan kalau tes dan butir-butir soal yang menjadi dasar penafsiran dan penyimpulan tersebut mempunyai kualitas rendah. Dengan demikian tujuan dikembangkannya sebuah tes untuk mengetahui kadar atau kualitas suatu atribut psikologis tidak bisa direalisasikan secara memuaskan.

Agar bisa mengembangkan butir soal dan tes yang berkualitas baik, diperlukan adanya pemahaman terhadap sifat dasar pengukuran atribut psikologis, butir soal dan tes dan aturan-aturan yang terkait dengan pengembangan butir soal dan tes di samping penguasaan yang baik terhadap bidang atau materi yang diukur melalui butir soal dan tes tersebut. Meskipun ini tidak berarti seseorang harus menjadi ahli psikometri agar dapat menghasilkan butir soal dan tes yang berkualitas, pemahaman dasar minimal terhadap asumsi-asumsi teoritis dan aturan-aturan dasaradalah sebuah keniscayaan bagi seorang yang terlibat dalam prosespenulisan dan pengembanganalat ukur atribut psikologis.

4. Pengukuran atributpsikologis

Sebuah tes dengan butir-butir soal yang membentuknya diciptakan untuk mengukur kualitas-kualitas psikologis seperti *penguasaan bahasa Arab* atau, lebih sempit,*kemampuan berbicara dalam bahasa Arab*.Istilah kualitas psikologis ini dalam bahasa

[9]Lyman, Howard B., *Test Scores and What They Mean* (6th ed.), (Boston: Allyn and Bacon, 1998), hal. 47

Inggris disebut dengan *construct*atau *psychological attribute*. Construct sering dibandingkan dengan kualitas fisik seperti tinggi, berat, dan warna kulit. Construct dapat didefinisikan sebagai konsep rekaan (hipotetis) yang merupakan produk dari imajinasi ilmiah seorang ilmuan yang berupaya mengembangkan teori-teori untuk menjelaskan perilaku manusia.[10]Menurut Sumadi Suryabrata, sebagai atribut psikologis, construct tidak seperti objek yang mempunyai wujud fisik yang dapat diinderai sehinggadapat diketahui kadar dan kualitasnyasecara lebih akurat. Pengkajian terhadap construct hanya dapat dilakukan secara tidak langsung melalui fenomena-fenomena fisik atau perilaku nyata yang dianggap sebagai manifestasi dari construct tersebut.[11]

Karena wujudnya yang tidak riil dan bahkan tidak dapat dikonfirmasi secara pasti, mengukur construct menjadi sangat berbeda dengan mengukur kualitas fisik seperti tinggi atau berat suatu benda. Perbedaannya tidak saja pada sifat pengukuran construct yang hanya dapat dilakukan secara tidak langsung, tetapi juga pada argumentasi eksistensi construct, kesahihan hubungan construct dengan perilaku nyataatau fenomena yang dijadikan indikator (*operational correspondence*), kompleksitas pengembangan alat ukur atau tes, serta penafsiran dan penyimpulan berdasarkan hasil pengukuran tersebut.Karena itu pengukuran construct menjadi sebuah proses yang kompleks, mulai dari perencanaan, pengembangan, pelaksanaan hingga penafsiran terhadap informasi yang diperoleh mengenai construct tersebut. Disamping itu ancaman kesesatan pengukuran, baik sistematis maupun acak, merupakan bagian yang inheren dalam setiap pengukuran construct.

C. Penjaminan Kualitas Butir Soal

Butir-butir soal sebagai bahan dasar yang membangun sebuah tes dapat diumpamakan seperti sel genetik yang menentukan postur dan ciri suatu organisme. Oleh karena itu pentingnya menulis butir soal yang berkualitas baik telah sejak

---

[10]Crocker,Linda& Algina, *James, Introduction to Classical …,* hal. 4

[11] Sumadi Suryabrata, *Pengembangan Alat Ukur Psikologis,* (Yogyakarta: Penerbit Andi, 2000), hal. 17

lama disampaikan. Cronbach, salah seorang ahli psikometri terkemuka, pada tahun 70an telah mengingatkan perlunya sebuah teori yang dapat dijadikan dasar dan rujukan bagi para penulis butir soal dan tes sehingga kegiatan tersebut menjadi terstandar.[12] Namun hingga akhir abad kedua puluh, teori yang didambakan tersebut belum dapat diwujudkan. Haladyna[13] dan Osterlind[14] masih mengeluhkan apa yang dikeluhkan oleh Cronbach sekitar 30 tahun yang lalu atas tidak adanya sebuah teori tentang penulisan butir soal. Sementara bagian-bagian lain dari disiplin ilmu pengukuran mental telah menjalani kemajuan pesat, masalah penulisan butir soal tidak mengalami perkembangan seperti yang diharapkan.

Meskipun demikian berbagai upaya dan pemikiran yang disampaikan oleh para ahli selama kurun waktu setengah abad tersebut dapat dijadikan pedoman untuk menghasilkan butir-butir soal yang berkualitas. Sembari mengupayakan sebuah solusi yang lebih mendasar berupa sebuah teori untuk pengembangan dan penulisan butir soal, pemikiran-pemikiran tersebut dapat dijadikan rujukan. Penulis berkeyakinan dengan mengikuti panduan-panduan yang ada sekarang ini saja seperti yang telah disusun para ahli, akan dapat dicapai peningkatan kualitas butir soal dan tes yang signifikan.

Pada pemaparan di bawah ini akan dibahas kriteria-kriteria yang harus dimiliki oleh butir soal yang baik dan karena itu penulis soal harus mengupayakan butir-butir soal yang ditulisnya mencapai tingkatan tersebut. Sebelum dipaparkan kriteria butir soal yang baik yang harus menjadi standar dan target capaian bagi penulis soal, terlebih dahulu akan dibicarakan pengertian dari butir soal itu sendiri.

1. Pengertian butir soal

Butir soal, dan tes, merupakan ukuran yang digunakan untuk mengukur construct. Melalui butir soal dan tes, construct diterjemahkan ke dalam bentuk operasional dalam bentuk perilaku yang diharapkan ditampilkan oleh orang (testee) yang merespon

---

[12]Haladyna, Thomas M., *Developing and Validating …* , hal.vii

[13]Ibid. , hal. vii

[14]Osterlind,Steven J., *Constructing Test …,* hal. 1 - 4

kepada stimulus yang disampaikan melalui tes. Perilaku yang ditampilkan oleh testee merupakan respon mereka terhadap stimulus (butir soal dan tes) yang dipaparkan kepada mereka. Dari respon tersebut akan disimpulkan kadar dan kualitas construct yang dimiliki oleh testee yang menjawab tes tersebut.

Sejalan dengan informasi tentang kaitan butir soal dan tes di atas, Osterlind mengemukakan pengertian butir soal dalam pemeriksaan atribut psikologis sebagai sebuah satuan pengukuran dengan sebuah stimulus dan ketentuan untuk menjawabnya dimana butir soal tersebut dimaksudkan untuk menghasilkan respon dari testee sehingga kinerja dalam suatu construct dapat diketahui.[15]

Dalam prakteknya sebuah butir soal saja tidaklah memadai untuk mengukur suatu constuct karena atribut psikologis adalah sesuatu yang sangat kompleks. Untuk meminimalisir dan mengkompensasi kesesatan pengukuran serta mencakup secara representatif berbagai dimensi yang terdapat dalam sebuah construct, diperlukan butir-butir soal dalam jumlah yang memadai. Kumpulan butir-butir soal yang mengukur sebuah construct membentuk tes. Crocker dan Algina mendefinisikan tes sebagai sebuah prosedur baku untuk memperoleh sampel perilaku dari domain atau construct tertentu.[16]

Bentuk atau format butir soal pada masa sekarang ini sangat beragam. Secara garis besar format butir soal terbagi kepada tiga macam. Yang pertama adalah butir soal dengan pilihan-pilihan jawaban yang disediakan, dimana testee diminta memilih jawaban yang benar atau paling benar dari pilihan-pilihan tersebut. Jenis soal yang pertama ini termasuk pilihan ganda, menjodohkan dan benar-salah. Yang kedua adalah *constructed-response*(tanggapan yang dikonstruksikan) dimana testee harus memikirkan dan memformulasikan sendiri jawaban yang benar dan menuliskan pada tempat yang ditentukan. Jenis yang kedua ini termasuk tes jawaban singkat, mengisi yang kosong dan esei. Yang ketiga adalah tes yang berbentuk kinerja dimana testee diminta untuk menampilkan suatu perbuatan yang menggambarkan suatu kemampuan atau ketrampilan untuk dinilai oleh pengamat.

---

[15] Osterlind, Steven J., *Constructing Test ... ,* hal. 19

[16]Crocker, Linda & Algina, James, *Introduction to Classical ...,* hal. 4

2. Kriteria butir soal yang baik

Menulis butir-butir soal yang baik merupakan pekerjaan yang kompleks. Pekerjaan ini membutuhkan tidak hanya pengetahuan tentang pengembangan alat ukur kognitif, tetapi juga rasa seni dan kreatifitas. Yang pertama, yaitu pengetahuan, dapat dipelajari dengan kesungguhan. Tetapi yang kedua, rasa seni dan kreatifitas hanya dapat diasah dengan latihan dan pengalaman.

Sebagian ahli seperti Haladyna membuat sebuah panduan penulisan butir-butir soal yang fokus pada pilihan ganda (*multiple choice*).[17]Penulis yang sama juga merumuskan panduan serupa untuk penulisan butir-butir soal yang mengukur kemampuan berfikir tingkat yang lebih tinggi (*higher order* thinking).[18]Gronlund, ahli yang lain, juga mencoba membuat pekerjaan menulis dan mengembangkan butir soal menjadi lebih mudah dan sederhana. Dia mengemukakan sejumlah aturan yang mencakup hal-hal sebaiknya dilakukan dan hal-hal yang sebaiknya dihindari dalam penulisan butir soal.[19]Demikian juga Popham menyajikan beberapa panduan untuk menghasilkan butir soal yang berkualitas.[20]Tetapi bagi ahli lain sebuah daftar "kerjakan" dan "jangan kerjakan" yang memandu penulisan butir soal tidak memadai, karena pendekatan yang seperti ini tidak mencerminkan kompleksitas dari pekerjaan tersebut. Oleh sebab itu sebuah daftar yang memuat aturan sederhana sulit untuk menghasilkan butir-butir soal yang berkualitas baik dengan kelemahan-kelemahan yang akan mengurangi validitas tes.

Untuk mengatasi hambatan tersebut, Osterlind mengemukakan tujuh kriteria yang lebih umum yang jika diikuti akan memastikan butir-butir soal yang ditulis akan memenuhi kriteria butir soal yang berkualitas baik.[21] Kriteria-kriteria tersebut adalah sebagai berikut.

---

[17]Haladyna, Thomas M., *Developing and Validating …* hal.77

[18]Haladyna, Thomas M, *Writing Test Items to Evaluate Higher Order Thinking*, (Boston: Allyn and Bacon, 1977) hal. 67 -91

[19] Gronlund, Norman E., *Assessment of Student Achievement* (6th Ed), (Boston: Allyn and Bacon, 1998), hal. 60 -106

[20] Popham, W. James, *Classroom Assessment: What Teachers Need to Know*, (Boston: Allyn and Bacon, 2002), hal. 126 - 143

[21]Osterlind, Steven J., *Constructing Test … ,* hal. 41 -43

Kriteria yang pertama, hendaklah ada derajat kesesuaian yang tinggi antara butir soal tertentu dengan tujuan umum dari tes atau construct psikologis yang hendak diukur oleh butir soal tersebut. Kriteria kedua adalah perlu adanya definisi yang jelas bagi atribut psikologis yang akan diukur. Ketika konsep yang diukur merupakan konsep yang sangat luas, definisi yang jelas tentang tujuan atau construct psikologis akan menghapus dan menghilangkan kekaburan dan meningkatkan peluang kesesuaian akan dicapai. Bila atribut psikologis sebuah tes dapat dirumuskan dengan jelas, maka identifikasi kumpulan butir-butir soal (*item pool*) untuk mengukur atribut psikologis tersebut akan dapat dilakukan dengan baik.

Kriteria ketiga untuk menghasilkan butir soal yang baik yaitu kontribusi setiap butir soal terhadap kesesatan pengukuran (*measurement error*) skor tes hendaklah ditekan seminimal mungkin. Kesesatan pengukuran, baik acak maupun sistematis, akan berpengaruh terhadap reliabilitas sebuah tes, dan reliabilitas merupakan persyaratan yang sangat penting bagi validitas.

Kriteria keempat mensyaratkan format tes harus sesuai dengan tujuan tes dimana tujuan yang sederhana secara umum menghendaki format tes yang sederhana dibandingkan dengan tujuan yang lebih kompleks.

Kriteria kelimamenetapkan bahwa butir soal yang baik harus memenuhi asumsi-asumsi teknis. Asumsi-asumsi tersebut adalah unidimensionalitas butir soal dan *local independence*. Secara sederhana unidimensionalitas butir soal bermakna bahwa jawaban peserta tes terhadap sebuah butir soal dapat dikaitkan dengan satu kemampuan. Misalnya, jawaban benar seorang murid terhadap sebuah butir soal bahasa Arab berasal dari kemampuan bahasa Arabnya, bukan kemampuan lain seperti pengetahuan linguistik misalnya. Karena itu dalam menulis butir soal, harus diusahakan agar setiap butir soal "menyasar" satu kemampuan saja.

Adapun *local independence* sebuah butir soal, secara praktis dapat difahami bahwa jawaban seorang peserta tes terhadap sebuah butir soal tidak dipengaruhi dan secara statistik tidak tergantung pada jawabannya terhadap butir soal lain yang manapun juga. Untuk menjelaskan poin ini, kadang-kadang dalam sebuah tes dijumpai sebuah butir soal yang mengandung jawaban secara jelas atau tersirat untuk butir soal yang lain. Sebagian peserta tes yang

teliti akan menemukan jawaban "bantuan" ini sehingga dia akan dapat menjawab benar butir soal lain yang jawaban terdapat dalam butir soal tersebut walaupun pada awal peserta tes tersebut tidak mengetahui jawaban yang benar.

Kriteria keenam untuk tes yang baik adalah penulisan butir soal harus mengikuti standar penulisan yang baik yang mencakup tata bahasa, ejaan, tanda baca dan aturan kebahasaan lainnya. Adapun kriteria ketujuh adalah penulisan butir soal harus memenuhi ketentuan hukum dan etika. Kadang-kadang penulis tes cenderung untuk menempuh jalan yang mudah dengan mengambil butir-butir soal atau tes yang dikembangkan oleh orang atau lembaga lain tanpa mendapatkan izin dari orang atau lembaga tersebut. Tindakan ini termasuk kategori plagiasi yang tidak saja merupakan pelanggaran etika, tetapi juga pelanggaran hukum.

3. Mendokumentasi langkah-langkah pengembangan butir soal

Dalam proses pengembangan dan penulisan butir soal, langkah-langkah yang diambil atau dilalui perlu didokumentasikan. Setidaknya ada dua alasan mengapa dokumentasi langkah-langkah tersebut perlu dilakukan seperti yang dikemukakan Osterlind.[22] Yang pertama, pengembangan butir-butir soal sama dengan kegiatan keilmuan lainnya dimana suatu kegiatan dapat direplikasi oleh orang atau ilmuan yang lain. Bila prosedur seperti yang termuat dalam dokumen diikuti dengan benar, maka hasil yang dicapai kemungkinan besar akan sama, dalam fluktuasi peluang, dengan hasil aslinya.

Yang kedua, dokumentasi diperlukan karena deskripsi-deskripsi tujuan yang tercantum dalam dokumen tersebut akan membantu saat menentukan apakah penafsiran tertentu terhadap skor-skor tes adalah valid.Bahwa dokumen tersebut dapata membantu penafsiran dapat dipahami dengan jelas karena setiap butir soal dalam tes pasti ditulis untuk tujuan tertentu. Semakin jelas pernyataan tujuannya, maka akan semakin baik penafsiran yang dapat dilakukan.

Demikianlah validitas tes tidak dapat dipisahkan dari butir-butir soal yang darinya skor-skor tes diperoleh. Butir-butir soal yang berkualitas baik yang dikembangkan dengan proses dan

[22]Osterlind, Steven J., *Constructing Test ... ,* hal. 66

langkah-langkah yang dapat dipertanggungjawabkan akan menjadi fondasi yang kuat bagi penafsiran tes dengan derajat validitas yang tinggi.

D. Penutup

Sebuah tes dilaksanakan dengan ujuan untuk mengetahui kadar atribut psikologis tertentu pada testee sehingga kesimpulan yang berbasis bukti dapat diformulasikan dan tindakan yang relevan dapat dilaksanakan. Kedua fungsi itu tidak bisa dilakukan begitu saja dengan anggapan bahwa kesimpulan yang diambil dan tindakan yang dieksekusi adalah benar. Untuk melihat sejauh mana kebenaran keduanya perlu dilakukan penilaian bukti bukti pendukung, dasar pikir teoritis dan argumen argumen yang mendasarinya. Penilaian ini dikenal dengan istilah validitas tes.

Dalam kegiatan menilai validitas sebuah tes, skor tes merupakan data utama yang dijadikan dasar pijakan penafsiran. Agar kegiatan penilaian validitas itu menghasilkan deskripsi validitas yang mendekati kebenaran, skor tes yang menjadi objek analisanya haruslah dapat diyakini mencerminkan construct yang diukur dan merepresentasikan kadar construct dari masing masing testee. Kualitas tersebut hanya mungkin dicapai jika butir butir soal memenuhi kriteria butir soal yang baik.

Pengembangan dan penulisan butir soal yang baik adalah pekerjaan yang menuntut berbagai kecakapan dan keahlian pada diri orang yang melaksanakan tugas tersebut. Disamping itu, penulisan butir soal menuntut kreatifitas yang tinggi sehingga butir-butir soal yang dihasilkan tidak hanya memenuhi prinsip prinsip dan asumsi asumsi teoritis dan teknis psikometri, tetapi juga memiliki keindahan dan keragaman yang menarik testee yang menjawabnya.

DAFTAR PUSTAKA

Crocker, Linda & Algina, James,*Introduction to Classical and Modern Test Theory*, New York: CBS College Publishing, 1986

Gronlund, Norman E., *Assessment of Student Achievement*, Boston: Allyn and Bacon, 1998

Haladyna, Thomas M., *Developing and Validating Multiple-Choice Test Items* (2nd Ed), New Jersey: Lawrence ErlbaumAssociates, 1999

Haladyna, Thomas M, *Writing Test Items to Evaluate Higher Order Thinking*, Boston: Allyn and Bacon, 1977

Linn, R. L. *Educational Measurement* (3rd Ed), New York: Macmillan, 1988

Lyman, Howard B., *Test Scores and What They Mean* (6th ed.), Boston: Allyn and Bacon, 1998

Osterlind, Steven J., *Constructing Test Items: Multiple-Choice, Constructed Response, Performance and Other Formats*, (2nd Ed.), Boston: Kluwer Academic Publishers, 1998

Popham, W. James, *Classroom Assessment: What Teachers Need to Know*, Boston: Allyn and Bacon, 2002

Suryabrata, Sumadi, *Pengembangan Alat Ukur Psikologis*, Yogyakarta: Penerbit Andi, 2000